The Sigmoid Function

The sigmoid function of logistic regression may sometimes seem like an ad hoc assumption. However, under a generative exponential family distribution, the posterior over the classes naturally follows a sigmoid function.

The Exponential Family

Distributions as diverse as the Gaussian, Bernoulli, and Poisson are all examples of a more general class of distributions that form the exponential family. In order to qualify to be a member of this group, the distribution of the random variable x must, for some set of parameters η follow the form

$$p(x) = b(x)e^{\eta^T \Phi(x) - a(\eta)}$$

Here $\Phi(x)$ is some vector valued function of x that in our case may correspond to the features we're using to encode the independent variable x.

The Sigmoid Function

Suppose that we want to classify the points x into two classes $\mathcal{C}_1$ and $\mathcal{C}_2$ and suppose furthermore that the class conditional distributions of x follows a distribution from the exponential family. While these two distributions must be of the same type, they may nonetheless have different parameters (in fact, if they didn't there would be no difference between the two classes).

$$p(x|\mathcal{C}_1) = b(x)e^{\eta^T \Phi(x) - a(\eta)} \qquad p(x|\mathcal{C}_2) = b(x)e^{\rho^T \Phi(x) - a(\rho)}$$

Using Bayes' rule we can then compute the posterior distributions over the classes

$$p(\mathcal{C}_1|x) = \frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x)} = \frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x|\mathcal{C}_1)p(\mathcal{C}_1) + p(x|\mathcal{C}_2)p(\mathcal{C}_2)} = \frac{1}{1 + \frac{p(x|\mathcal{C}_2)p(\mathcal{C}_2)}{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}}$$

Using the form for distributions in the exponential family, we find

$$p(\mathcal{C}_1|x) = \frac{1}{1 + e^{(\rho-\eta)^T \Phi(x) + a(\eta) - a(\rho)} \cdot \frac{1-q}{q}}$$

where we have introduced the short hand $q = p(\mathcal{C}_1)$. Now, define

$$\theta_0 := a(\rho) - a(\eta) + \ln(q/(1-q)), \qquad \theta_i = \eta_i - \rho_i \qquad \text{for } i > 0.$$

Then we can write the posterior distribution above as (defining $\Phi(x)_0 = 1$)

$$p(\mathcal{C}_1|x) = \frac{1}{1 + e^{-\theta^T \Phi(x)}}$$

which of course is nothing but the sigmoid function. In other words, as long as the class conditional distributions belongs to the exponential family, the posterior distribution is a sigmoid. So, the sigmoid is not just pulled out of thin air but instead follows as the posterior distribution in a very general class of models.